scientific data

OPEN

DATA DESCRIPTOR

PCMR: a comprehensive precancerous molecular resource

Yichun Xiong[1,2], Jiaqi Li[1,2], Wang Jin[1], Xiaoran Sheng[1], Hui Peng[1], Zhiyi Wang[1], Caifeng Jia[1], Lili Zhuo[1], Yibo Zhang[1], Jingzhe Huang[1], Modi Zhai[1], Beibei Lyu[1], Jie Sun[1] ✉ & Meng Zhou[1] ✉

Early detection and intervention of precancerous lesions are crucial in reducing cancer morbidity and mortality. Comprehensive analysis of genomic, transcriptomic, proteomic and epigenomic alterations can provide insights into the early stages of carcinogenesis. However, the lacke of an integrated, well-curated data resource of molecular signatures limits our understanding of precancerous processes. Here, we introduce a comprehensive PreCancerous Molecular Resource (PCMR), which compiles 25,828 molecular profiles of precancerous samples paired with normal or malignant counterparts. These profiles cover precancerous lesions of 35 cancer types across 20 organs and tissues, derived from tissue samples, liquid biopsies, cell lines and organoids, with data from transcriptomics, proteomics and epigenomics. PCMR includes 62,566 precancer-gene associations derived from differential analysis and text-mining using the ChatGPT large language model. We examined PCMR dataset reliability and significance by the authoritative precancerous molecular signature, along with its biological and clinical relevance. Overall, PCMR will serve as a valuable resource for advancing precancer research and ultimately improving patient outcomes.

Background & Summary

Cancer remains one of the most significant global health challenges, despite significant advances in early detection, screening and treatment strategies[1]. Cancer progression is a complex, multistep process that typically begins with the transformation of normal cells into precancerous lesions, which may subsequently progress to invasive malignancies[2–4]. These transitions are characterized by abnormal histological and immunohistological features, as well as changes at the molecular level[5–12]. Early detection and intervention of precancerous lesions are critical for reducing cancer morbidity and mortality[13–18]. Understanding the molecular alterations that drive these transitions is crucial for developing effective early detection methods and therapeutic strategies. However, despite significant advances in the molecular profiling of cancer, there is still a critical gap in our ability to effectively detect and monitor pre-cancerous conditions.

High-throughput omics technologies, such as genomics, transcriptomics, proteomics, and epigenomics, have greatly advanced our understanding of precancerous lesions[19,20]. Numerous molecular profiling studies have revealed a large number of molecular alterations associated with tumorigenesis[19,21,22]. Despite these valuable contributions, significant gaps remain in the comprehensive monitoring and analysis of precancerous lesions[17,19,23]. Existing molecular profiling and knowledge of precancerous lesions is fragmented and scattered across various studies, publications, databases and research groups[24], making it challenging to integrate and compare findings. In addition, many existing cancer databases focus primarily on invasive malignancies, neglecting the earlier, non-invasive stages of cancer progression[25–29]. This data fragmentation hinders the discovery of common molecular signatures that may be present in precancerous lesions across different cancer types. Therefore, there is a critical need for a comprehensive, centralized resource that consolidates and harmonizes multi-omics data, specifically focused on precancerous lesions.

In this study, we present the Precancerous Molecular Resource (PCMR), the first comprehensive database specifically designed to consolidate and harmonize multi-omics data focused on precancerous lesions. PCMR integrates precancerous molecular profiles, gene-precancerous lesion associations, and functional modules to provide a comprehensive online platform for data retrieval, analysis and visualization (Fig. 1). The PCMR resource incorporates two main categories of data (Table 1): (1) High-throughput multi-omics profiles (Fig. 2A), including data from precancerous lesions and paired normal and/or malignant conditions, generated from various biological materials (e.g., tissue samples, liquid biopsies, cell lines, and organoids). The integration of

[1] School of Biomedical Engineering, Eye Hospital, Wenzhou Medical University, Wenzhou, 325027, P. R. China. [2] These authors contributed equally: Yichun Xiong, Jiaqi Li. ✉e-mail: suncarajie@wmu.edu.cn; zhoumeng@wmu.edu.cn

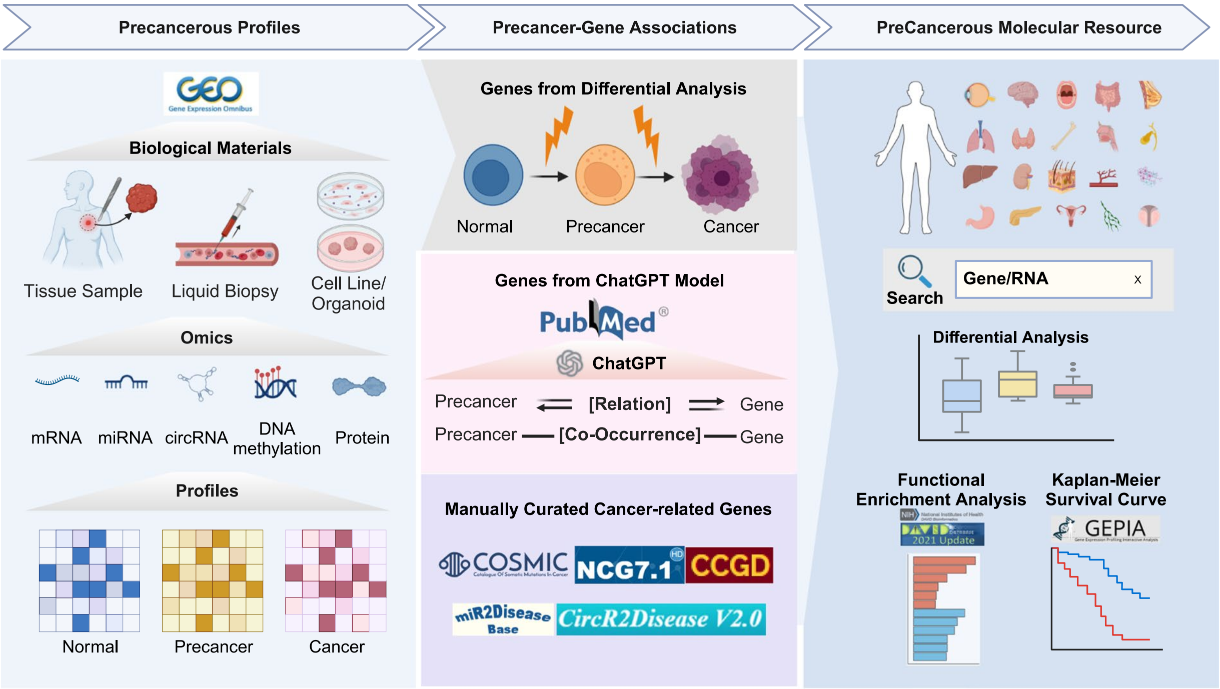


Fig. 1 Data source and architecture of PCMR (PreCancerous Molecular Resource). PCMR combines precancerous profiles, precancer-gene associations, and functional modules, providing a holistic system for online retrieval, analysis, and visualization of gene signature. The PCMR platform incorporates: (1) high-throughput profiles of precancer and paired normal and/or malignant conditions across various omics, platforms and biological material sources. (2) precancer-gene associations derived from differential analysis of precancerous profiles, text mining of abstracts using the ChatGPT large language model, and manual curation of cancer-related genes from published resources. (3) abundant functionality on the 'Search', 'Browser' and 'Analysis' pages, facilitating the retrieval, analysis and visualization of resources related to precancerous lesions.

	Total No.	Items	No. of each item
Precancerous profiles	25,828	mRNA	6,353
		microRNA	18,485
		circRNA	78
		DNA methylation	852
		Protein	60
Precancer-gene associations	62,566	Precancer-genes from differential analysis	17,230
		Precancer-genes from ChatGPT model	41,862
		Manually curated cancer-genes	3,474

Table 1. Precancerous profiles and precancer-gene associations in PCMR.

transcriptomic, proteomic, and epigenomic data, including mRNA, miRNA, circRNA, protein, and DNA methylation, allows for a more holistic exploration of precancerous lesions; (2) Precancer-gene associations, derived from differential analysis of precancerous profiles (Fig. 2B), text mining of abstracts using the ChatGPT large language model, and manual curation of cancer-related genes from published resources (Fig. 2C). In addition, PCMR provides abundant functionality on the 'Search', 'Browser' and 'Analysis' pages, enabling users to retrieve, analyze and visualize resources related to precancerous lesions. PCMR will serve as an essential resource for advancing our understanding of the cellular and biological processes underlying the onset and progression of precancer.

Methods

Precancerous profiles collection and quality control. To systematically collect high-throughput sequencing and microarray data from precancerous, paired normal and/or malignant conditions, we searched the Gene Expression Omnibus (GEO)[24] database with the following keywords: 'precancerous', 'premalignant', 'preneoplastic', 'preinvasive', 'precarcinoma lesion', 'cancer/tumor precursor', 'benign lesion', 'incipient neoplasia' and 'dysplasia'. Relevant data matrices for various premalignant lesions, including transcriptomic, epigenomic, and proteomic profiles, were manually curated and downloaded.

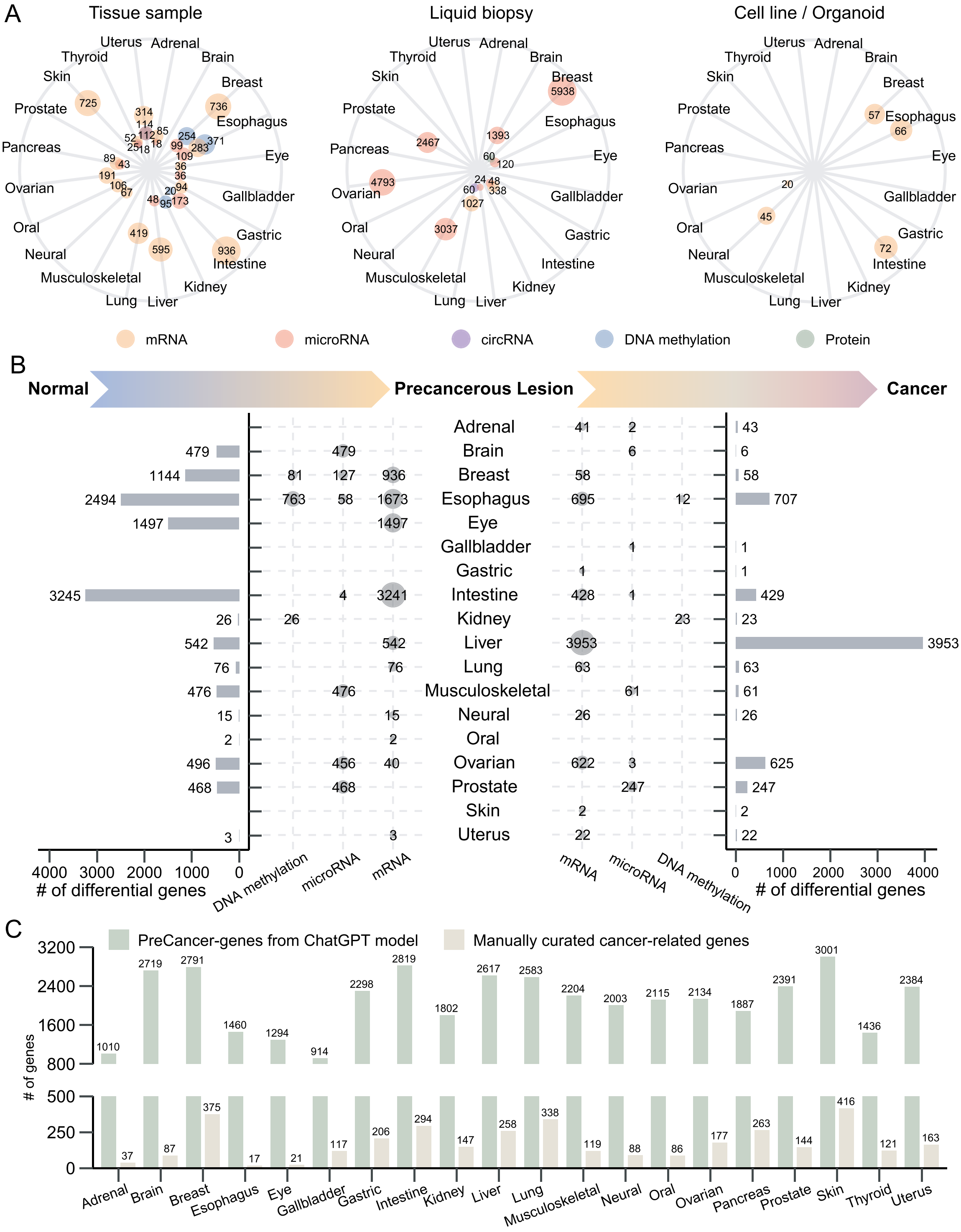


Fig. 2 Summary of precancerous profiles and precancer-gene associations in PCMR. (**A**) Number of precancerous profiles for each organ and tissue type, sourced from tissue biopsies, liquid biopsies, cell lines and organoids, encompassing omics data types including mRNA, microRNA, circRNA, DNA methylation and protein. (**B**) Precancer-gene associations identified from differential analysis were summarized. For mRNA, microRNA, circRNA, and protein analysis, a fold change greater than 2 and an adjusted *P*-value below 0.05 were used to identify differential genes. In the differential DNA methylation analysis, genes with an absolute methylation difference greater than 0.2 and an adjusted *P*-value less than 0.05 were considered differentially methylated. (**C**) Precancer-gene associations from text mining of literature abstracts using the ChatGPT large language model, and manual curation of cancer-related genes from five published resources across cancer types.

To ensure the quality and reliability of the data included in our repository, we implemented a rigorous multi-step filtering process. We first identified datasets using keywords related to precancerous lesions. The following criteria were used to exclude datasets: 1) datasets involving physical, chemical or biological interventions (e.g., viral infection, drug treatment, radiation exposure) that could introduce confounding effects; 2) datasets lacking precancerous lesion samples or missing both normal and cancer samples; 3) datasets without molecular omics data or sufficient annotation details; 4) datasets with incomplete or ambiguous clinical information regarding sample disease status; 5) datasets with fewer than 20 total samples; 6) datasets with the sample group (precancerous, normal/cancer) contained fewer than 3 samples; 7) datasets with duplicate dataset IDs and sample IDs to ensure unique storage and analysis.

The following criteria were used to filter samples and genes: 1) samples containing mixed tissues, where precancerous lesions and cancerous tissues could not be clearly distinguished were excluded; 2) samples with uncertain disease classification were removed; 3) genes or RNAs that could not be mapped to standardized names or IDs were excluded; 4) genes or RNAs detected in less than 20% of samples within a dataset were removed to increase data reliability.

After filtering, datasets were manually curated and cross-checked for consistency before being stored and analyzed. We manually reviewed the descriptions and literature of each dataset following keyword searches in GEO. For sample annotations, including patient clinical information, data platform annotations, and source information, we implemented a two-step verification process: independent data collection followed by cross-checking. The dataset and sample information were cross-checked by different researchers against the source databases and original publications. Any discrepancies identified during cross-checking were systematically reviewed, and corrections were made to ensure their consistency with the original source. In cases where uncertainty remained, discussions were held among the curators to reach a consensus, and a final decision was made based on the most reliable source documentation. These stringent quality control measures ensure that only high-quality, well-annotated datasets are included in our repository.

Data normalization.

After completing the above steps, data normalization was performed. For each profile, sample IDs were unified into GEO accession IDs. Gene IDs in RNA and protein datasets were unified into Gene Symbols, microRNA IDs were aligned with miRBase precursor accession IDs, and circRNA IDs were converted to circBase IDs. For methylation datasets, methylation profiles were normalized to the gene level, defined as the average methylation level of probes within the promoter region (2 kb upstream to 0.5 kb downstream of the transcription start site).

Normalization was performed across arrays for each microarray dataset. High-throughput RNA-sequencing profiles were saved as either FPKM (Fragments Per Kilobase of transcript per Million mapped reads) or RPKM (reads per kilobase of transcript per million reads mapped). The human reference genome GRCh38 was used in all datasets.

Precancerous lesion-gene associations from differential analysis.

In PCMR, genes relevant to precancerous lesions were primarily identified through differential analysis and text-mining using the ChatGPT large language model. Differential analysis was performed independently for precancerous samples against their paired normal or cancer counterparts in each dataset. Significantly differential genes associated with precancerous lesions were identified using thresholds for fold change, absolute difference, and adjusted *P*-values. Differential analyses were conducted by R (version 4.2.0), using packages including 'rstatix', 'dplyr', 'tidyr' and 'limma'[30]. The human reference genome GRCh38 was used for all datasets and visualization. The *P*-value for comparing two groups in differential analysis was calculated using the Mann-Whitney U test and adjusted with the Bonferroni correction. In differential DNA methylation analysis, genes with an absolute methylation difference greater than 0.2 and an adjusted *P*-value less than 0.05 were considered differentially methylated. For mRNA, microRNA, circRNA, and protein analysis, a fold change greater than 2 and an adjusted *P*-value below 0.05 were considered as differential.

Precancerous lesion-gene associations from literatures.

For precancer-gene associations derived through text mining, gene names, precancerous lesion names, and human organ/tissue names were extracted from the abstracts and unified. Initially, we searched the PubMed database using the same keywords as for high-throughput profiling, which yielded 717,875 article abstracts from relevant studies. We then used ChatGPT's standard GPT-3.5-turbo model (unmodified by fine-tuning) to process the abstracts, implementing a custom prompt template for biomedical entity recognition via the Azure OpenAI API. The prompt is "You are a professional biologist. Your task is to accurately identify and list the following categories from the given abstract: organ, gene symbol, precancerous lesion, sentences containing both gene symbols and precancerous lesions. Organs should be categorized into given list (organ list)". After initially filtering out articles with missing organ or gene information, we narrowed down the dataset to 147,110 articles, which were then subjected to a second round of information extraction. Finally, we obtained meaningful precancerous gene information from 94,888 articles using the ChatGPT platform.

Since our goal was to extract genes associated with precancerous lesions across various human organs and tissues, we did not impose strict criteria on lesion subtypes or anatomical locations within an organ. Instead, we focused on extracting and unifying human organ/tissue names and gene names. During the data mining process with ChatGPT, we first standardized organ and tissue nomenclature to ensure consistency. This standardized terminology was used as a reference in the prompts to guide ChatGPT in categorizing extracted terms appropriately. After ChatGPT extracted human organ/tissue names, gene names, and precancerous lesion names from the abstracts, we retained entries where both an organ/tissue and a gene name were present, without requiring specific precancerous lesion names. For gene name standardization, we removed rows containing empty or meaningless strings in the gene column, and then used the Ensembl, miRBase, and circBase databases to ensure uniformity in gene symbols and IDs. Precancerous lesion names were converted to lowercase and stripped of unnecessary or meaningless characters before storage. Finally, the precancer-gene associations were categorized into two confidence levels: those inferred to be related contextually, and those identified through co-occurrence within the same abstract.

Additionally, PCMR also incorporates cancer-associated genes across various cancer types from five publicly available, manually curated databases, including the Catalogue Of Somatic Mutations In Cancer (COSMIC)[31], the Network of Cancer Genes (NCG)[32], the Candidate Cancer Gene Database (CCGD)[33], miR2Disease[34], CircR2Disease[35].

Data Records

The dataset is available at Figshare[36] and the dataset contains:

(i) The 'precancerous profiles information.txt' file contains detailed information of precancerous profiles paired with normal and/or malignant counterparts. The table includes organ/tissue, cancer type, disease state (normal/premalignant/cancer), disease name, biological materials, omics, platform, gender, age, cancer stage, and PubMed ID.

(ii) The 'precancer-gene associations from differential analysis.txt' file contains precancer-gene associations derived from differential analysis. The table includes organ/tissue, gene symbol, cancer type, biological material origin, omics, and the relevant differential groups.

(iii) The 'precancer-gene associations from ChatGPT.txt' file contains precancer-gene associations obtained from the ChatGPT large language model. The table includes organ/tissue, gene symbol, cancer type, PubMed ID, and whether the association was obtained from relation extraction.

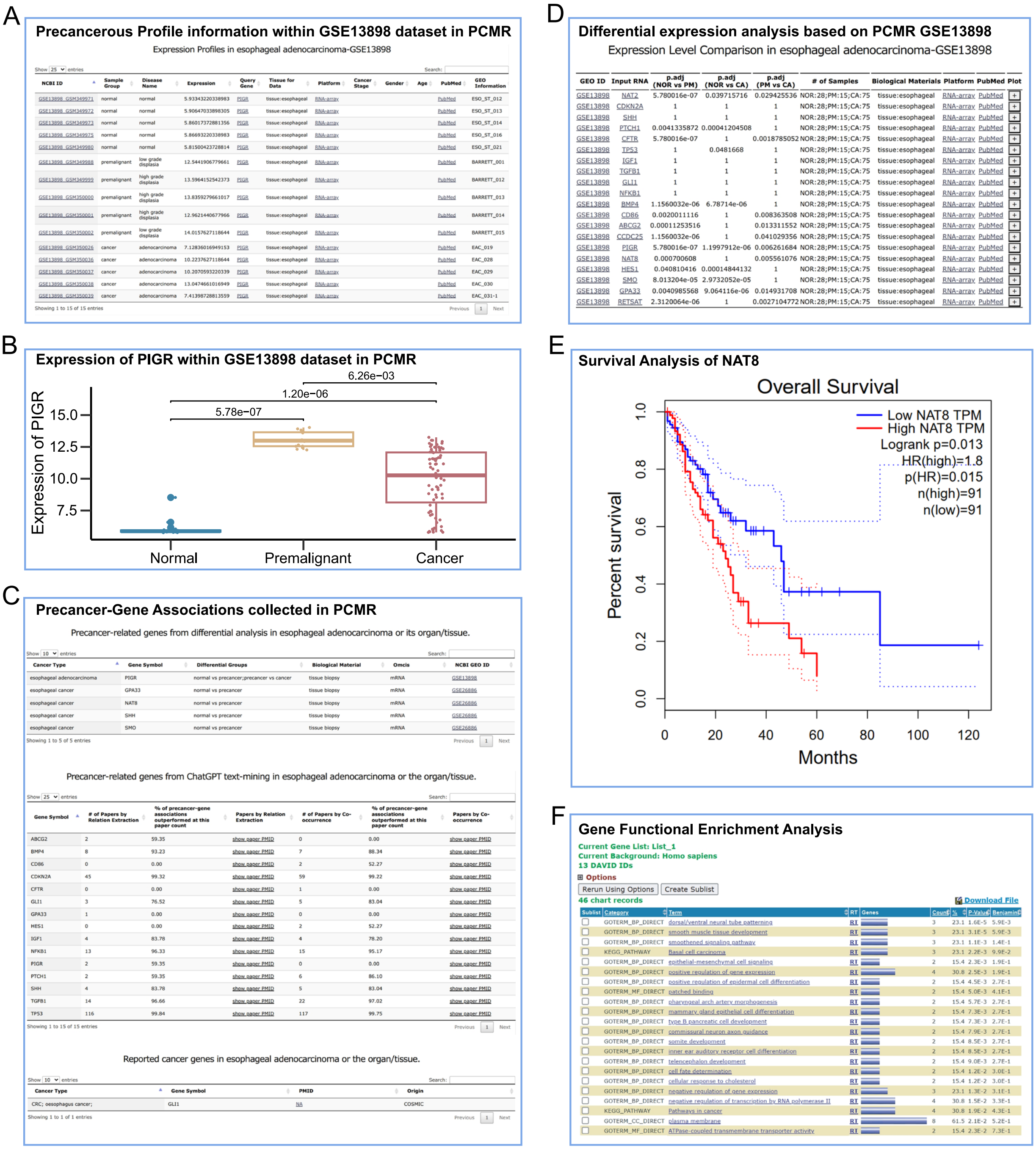
(iv) The 'cancer-gene associations from manually curated databases.txt' file contains cancer-associated genes or RNAs from five publicly available, manually curated databases, including COSMIC, NCG, CCGD, miR2Disease and CircR2Disease. The table includes organ/tissue, gene symbol, cancer type, PubMed ID, data origin.

(v) The 'Technical Validation data.xlsx' file contains all the data used during technical validation of PCMR. The file includes 1) Precancerous sample clinical information and profile metadata of GSE13898 dataset cataloged in PCMR. 2) PIGR gene expression level across precancerous, normal, and/or malignant samples within GSE13898. 3) Precancer-gene associations collected in PCMR, derived from differential analysis of precancerous profiles, text mining of abstracts using the ChatGPT large language model, and manual curation of cancer-related genes from published resources. 4) Differential expression analysis results and expression level of 7 previously reported esophageal precancer-related genes (*CFTR*, *PIGR*, *NAT2*, *CCDC25*, *ABCG2*, *CD86*, *GPA33*), and 13 additional unreported relevant genes (*HES1*, *RETSAT*, *NAT8*, *BMP4*, *SHH*, *TGFB1*, *SMO*, *CDKN2A*, *PTCH1*, *IGF1*, *GLI1*, *TP53*, *NFKB1*) from GSE13898 dataset within PCMR. 5) Overall survival data of esophageal carcinoma from The Cancer Genome Atlas Program (TCGA), together with *NAT8* expression level and its subgroups based on median. 6) Gene set functional items from gene functional enrichment analysis of 13 differentially expressed genes based on the Database for Annotation, Visualization and Integrated Discovery (DAVID).

Technical Validation

To ensure the accuracy of clinical information, multi-omics profiles and precancer-gene associations in PCMR, we followed a series of carefully designed procedures. For the collection of high-throughput sequencing and microarray profiles from precancerous, paired normal, and/or malignant conditions, we manually reviewed the descriptions and literature of each dataset after performing keyword searches in GEO. For annotation details of each sample such as patient clinical information, data platform annotations and data source information, we ensure the accuracy and consistency of the extracted information data by implementing a two-step verification process, including independent data collection and cross-checking. Collected dataset and sample information were then independently cross-checked by different researchers against the source databases and original publications. Any discrepancies identified during cross-checking were systematically reviewed, and corrections were made to ensure consistency with the original source information. In cases where uncertainty remained, discussions were held among the curators to reach a consensus, and a final decision was made based on the most reliable source documentation.

On the other hand, a strict and rigorous approach was followed for the retrieval of precancer gene associations. For associations derived from differential analysis of high-throughput molecular profiles, R packages ('rstatix', 'dplyr', 'tidyr' and 'limma') was used. The *P*-value for comparing two groups was calculated by the Mann-Whitney U test and adjusted with the Bonferroni correction. For mRNA, microRNA, circRNA, and protein analysis, a fold change greater than 2 and an adjusted *P*-value below 0.05 were considered as differential. In the differential DNA methylation analysis, genes with an absolute methylation difference greater than 0.2 and an adjusted *P*-value less than 0.05 were considered differentially methylated. By these commonly used differential analysis tools, the data underwent filtering and normalization, followed by the application of strict thresholds for each type of omics. The entire process was independently cross-verified for reliability. For associations extracted from text mining, we carefully selected ChatGPT prompts and pretested the model using a small paper dataset containing precancer-gene pairs. Cancer-related genes were collected from five widely used, manually curated public databases.



Precancerous Profile information within GSE13898 dataset in PCMR

Expression Profiles in esophageal adenocarcinoma-GSE13898

Show 25 entries Search:

NCBI ID	Sample Group	Disease Name	Expression	Query Gene	Tissue for Data	Platform	Cancer Stage	Gender	Age	PubMed	GEO Information
GSE13898_GSM349971	normal	normal	5.93343220338983	PIGR	tissue:esophageal	RNA-array				PubMed	ESO_ST_012
GSE13898_GSM349972	normal	normal	5.90647033898305	PIGR	tissue:esophageal	RNA-array				PubMed	ESO_ST_013
GSE13898_GSM349973	normal	normal	5.86017372881356	PIGR	tissue:esophageal	RNA-array				PubMed	ESO_ST_014
GSE13898_GSM349975	normal	normal	5.86693220338983	PIGR	tissue:esophageal	RNA-array				PubMed	ESO_ST_016
GSE13898_GSM349980	normal	normal	5.81500423728814	PIGR	tissue:esophageal	RNA-array				PubMed	ESO_ST_021
GSE13898_GSM349988	premalignant	low grade displasia	12.5441906779661	PIGR	tissue:esophageal	RNA-array				PubMed	BARRETT_001
GSE13898_GSM349999	premalignant	high grade displasia	13.5964152542373	PIGR	tissue:esophageal	RNA-array				PubMed	BARRETT_012
GSE13898_GSM350000	premalignant	high grade displasia	13.8359279661017	PIGR	tissue:esophageal	RNA-array				PubMed	BARRETT_013
GSE13898_GSM350001	premalignant	high grade displasia	12.9621440677966	PIGR	tissue:esophageal	RNA-array				PubMed	BARRETT_014
GSE13898_GSM350002	premalignant	low grade displasia	14.0157627118644	PIGR	tissue:esophageal	RNA-array				PubMed	BARRETT_015
GSE13898_GSM350026	cancer	adenocarcinoma	7.12836016949153	PIGR	tissue:esophageal	RNA-array				PubMed	EAC_019
GSE13898_GSM350036	cancer	adenocarcinoma	10.2237627118644	PIGR	tissue:esophageal	RNA-array				PubMed	EAC_028
GSE13898_GSM350037	cancer	adenocarcinoma	10.2070593220339	PIGR	tissue:esophageal	RNA-array				PubMed	EAC_029
GSE13898_GSM350038	cancer	adenocarcinoma	13.0474661016949	PIGR	tissue:esophageal	RNA-array				PubMed	EAC_030
GSE13898_GSM350039	cancer	adenocarcinoma	7.41398728813559	PIGR	tissue:esophageal	RNA-array				PubMed	EAC_031-1

Showing 1 to 15 of 15 entries Previous 1 Next

Expression of PIGR within GSE13898 dataset in PCMR

Precancer-Gene Associations collected in PCMR

Precancer-related genes from differential analysis in esophageal adenocarcinoma or its organ/tissue.

Show 10 entries Search:

Cancer Type	Gene Symbol	Differential Groups	Biological Material	Omcis	NCBI GEO ID
esophageal adenocarcinoma	PIGR	normal vs precancer;precancer vs cancer	tissue biopsy	mRNA	GSE13898
esophageal cancer	GPA33	normal vs precancer	tissue biopsy	mRNA	GSE26886
esophageal cancer	NAT8	normal vs precancer	tissue biopsy	mRNA	GSE26886
esophageal cancer	SHH	normal vs precancer	tissue biopsy	mRNA	GSE26886
esophageal cancer	SMO	normal vs precancer	tissue biopsy	mRNA	GSE26886

Showing 1 to 5 of 5 entries Previous 1 Next

Precancer-related genes from ChatGPT text-mining in esophageal adenocarcinoma or the organ/tissue.

Show 25 entries Search:

Gene Symbol	# of Papers by Relation Extraction	% of precancer-gene associations outperformed at this paper count	Papers by Relation Extraction	# of Papers by Co-occurrence	% of precancer-gene associations outperformed at this paper count	Papers by Co-occurrence
ABCG2	2	59.35	show paper PMID	0	0.00	show paper PMID
BMP4	8	93.23	show paper PMID	7	88.34	show paper PMID
CD86	0	0.00	show paper PMID	2	52.27	show paper PMID
CDKN2A	45	99.32	show paper PMID	59	99.22	show paper PMID
CFTR	0	0.00	show paper PMID	1	0.00	show paper PMID
GLI1	3	76.52	show paper PMID	5	83.04	show paper PMID
GPA33	1	0.00	show paper PMID	0	0.00	show paper PMID
HES1	0	0.00	show paper PMID	2	52.27	show paper PMID
IGF1	4	83.78	show paper PMID	4	78.20	show paper PMID
NFKB1	13	96.33	show paper PMID	15	95.17	show paper PMID
PIGR	2	59.35	show paper PMID	0	0.00	show paper PMID
PTCH1	2	59.35	show paper PMID	6	86.10	show paper PMID
SHH	4	83.78	show paper PMID	5	83.04	show paper PMID
TGFB1	14	96.66	show paper PMID	22	97.02	show paper PMID
TP53	116	99.84	show paper PMID	117	99.75	show paper PMID

Showing 1 to 15 of 15 entries Previous 1 Next

Reported cancer genes in esophageal adenocarcinoma or the organ/tissue.

Show 10 entries Search:

Cancer Type	Gene Symbol	PMID	Origin
CRC; oesophagus cancer;	GLI1	NA	COSMIC

Showing 1 to 1 of 1 entries Previous 1 Next

Differential expression analysis based on PCMR GSE13898

Expression Level Comparison in esophageal adenocarcinoma-GSE13898

GEO ID	Input RNA	p.adj (NOR vs PM)	p.adj (NOR vs CA)	p.adj (PM vs CA)	# of Samples	Biological Materials	Platform	PubMed	Plot
GSE13898	NAT2	5.780016e-07	0.039715716	0.029425536	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	CDKN2A	1	1	1	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	SHH	1	1	1	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	PTCH1	0.0041335872	0.00041204508	1	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	CFTR	5.780016e-07	1	0.0018785052	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	TP53	1	0.0481668	1	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	IGF1	1	1	1	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	TGFB1	1	1	1	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	GLI1	1	1	1	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	NFKB1	1	1	1	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	BMP4	1.1560032e-06	6.78714e-06	1	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	CD86	0.0020011116	1	0.008363508	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	ABCG2	0.00011253516	1	0.013311552	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	CCDC25	1.1560032e-06	1	0.041029356	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	PIGR	5.780016e-07	1.1997912e-06	0.006261684	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	NAT8	0.000700608	1	0.005561076	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	HES1	0.040810416	0.00014844132	1	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	SMO	8.013204e-05	2.9732052e-05	1	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	GPA33	0.0040985568	9.064116e-06	0.014931708	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+
GSE13898	RETSAT	2.3120064e-06	1	0.0027104772	NOR:28;PM:15;CA:75	tissue:esophageal	RNA-array	PubMed	+

Survival Analysis of NAT8

Gene Functional Enrichment Analysis

Current Gene List: List_1
Current Background: Homo sapiens
13 DAVID IDs
⊞ Options
Rerun Using Options Create Sublist
46 chart records Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_BP_DIRECT	dorsal/ventral neural tube patterning	RT		3	23.1	1.6E-5	5.9E-3
□	GOTERM_BP_DIRECT	smooth muscle tissue development	RT		3	23.1	3.1E-5	5.9E-3
□	GOTERM_BP_DIRECT	smoothened signaling pathway	RT		3	23.1	1.1E-3	1.4E-1
□	KEGG_PATHWAY	Basal cell carcinoma	RT		3	23.1	2.2E-3	9.9E-2
□	GOTERM_BP_DIRECT	epithelial-mesenchymal cell signaling	RT		2	15.4	2.3E-3	1.9E-1
□	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		4	30.8	2.5E-3	1.9E-1
□	GOTERM_BP_DIRECT	positive regulation of epidermal cell differentiation	RT		2	15.4	4.5E-3	2.7E-1
□	GOTERM_MF_DIRECT	patched binding	RT		2	15.4	5.0E-3	4.1E-1
□	GOTERM_BP_DIRECT	pharyngeal arch artery morphogenesis	RT		2	15.4	5.7E-3	2.7E-1
□	GOTERM_BP_DIRECT	mammary gland epithelial cell differentiation	RT		2	15.4	7.3E-3	2.7E-1
□	GOTERM_BP_DIRECT	type B pancreatic cell development	RT		2	15.4	7.3E-3	2.7E-1
□	GOTERM_BP_DIRECT	commissural neuron axon guidance	RT		2	15.4	7.9E-3	2.7E-1
□	GOTERM_BP_DIRECT	somite development	RT		2	15.4	8.5E-3	2.7E-1
□	GOTERM_BP_DIRECT	inner ear auditory receptor cell differentiation	RT		2	15.4	8.5E-3	2.7E-1
□	GOTERM_BP_DIRECT	telencephalon development	RT		2	15.4	9.0E-3	2.7E-1
□	GOTERM_BP_DIRECT	cell fate determination	RT		2	15.4	1.2E-2	3.0E-1
□	GOTERM_BP_DIRECT	cellular response to cholesterol	RT		2	15.4	1.2E-2	3.0E-1
□	GOTERM_BP_DIRECT	negative regulation of gene expression	RT		3	23.1	1.3E-2	3.1E-1
□	GOTERM_BP_DIRECT	negative regulation of transcription by RNA polymerase II	RT		4	30.8	1.5E-2	3.3E-1
□	KEGG_PATHWAY	Pathways in cancer	RT		4	30.8	1.9E-2	4.3E-1
□	GOTERM_CC_DIRECT	plasma membrane	RT		8	61.5	2.1E-2	5.2E-1
□	GOTERM_MF_DIRECT	ATPase-coupled transmembrane transporter activity	RT		2	15.4	2.3E-2	7.3E-1

Fig. 3 Technical validation of PCMR resource. (**A**) Precancerous sample clinical information and profile metadata of GSE13898 dataset cataloged in PCMR. PCMR incorporates information of each precancerous, normal and/or malignant sample, including sample group (normal/premalignant/cancer), disease name, biological tissue for data, data platform, cancer stage, gender, age, and PubMed ID. (**B**) Distribution of *PIGR* gene expression across precancerous, normal, and/or malignant conditions within GSE13898. *P*-value was calculated by Mann–Whitney U test, and adjusted by Bonferroni correction. (**C**) Precancer-gene associations collected in PCMR, originated from differential analysis of precancerous profiles, text mining of abstracts using the ChatGPT large language model, and manual curation of cancer-related genes from published resources. (**D**) Differential expression analysis of 7 previously reported esophageal precancer-related genes (*CFTR*, *PIGR*, *NAT2*, *CCDC25*, *ABCG2*, *CD86*, *GPA33*), and 13 additional unreported relevant genes (*HES1*, *RETSAT*, *NAT8*, *BMP4*, *SHH*, *TGFB1*, *SMO*, *CDKN2A*, *PTCH1*, *IGF1*, *GLI1*, *TP53*, *NFKB1*). Based on GSE13898 dataset, the expression level of RNAs are compared across precancerous, normal, and malignant conditions. *P*-value was calculated by Mann–Whitney U test, and adjusted by Bonferroni correction. (**E**) Kaplan-Meier survival analysis of *NAT8* in esophageal carcinoma based on the Gene Expression Profiling Interactive Analysis (GEPIA) database, with data from The Cancer Genome Atlas Program (TCGA). (F) Gene functional enrichment analysis of 13 differentially expressed genes based on the Database for Annotation, Visualization and Integrated Discovery (DAVID).

To verify the reliability of data resources and function included in PCMR, we examined the gene *PIGR* in the context of esophageal premalignant development, a relationship previously reported in the literatures[37–39]. Based on premalignant and normal/cancer samples from dataset GSE13898[40] stored within PCMR, we identified significant expression differences for *PIGR* between premalignant and normal/cancer samples (Fig. 3A, B). Furthermore, the association between esophageal precancer and *PIGR* derived from ChatGPT large language model was also collected in PCMR resource (Fig. 3C).

For a more thorough validation, we investigate the dynamic changes, biological processes, and clinical relevance of gene signatures within precancerous conditions, by including 7 previously reported esophageal precancer-related genes (*CFTR*[41,42], *PIGR*[37–39], *NAT2*[43,44], *CCDC25*[45], *ABCG2*[46–48], *CD86*[49], *GPA33*[50]), and 13 additional genes (*HES1*, *RETSAT*, *NAT8*, *BMP4*, *SHH*, *TGFB1*, *SMO*, *CDKN2A*, *PTCH1*, *IGF1*, *GLI1*, *TP53*, *NFKB1*). We observed that all seven previously reported genes showed significant differential expression between premalignant and normal/cancer samples based on dataset collected in PCMR (Fig. 3D). Among 13 genes with unreported relevant genes, 6 exhibited significant differential expression changes between precancerous lesions and normal or cancer groups (Fig. 3D). *RETSAT* and *NAT8* showed marked expression differences in both precancerous vs. normal and precancerous vs. cancer comparisons but no variation between cancer and normal samples, suggesting transient involvement in early carcinogenic processes. Conversely, *BMP4*, *SMO*, *PTCH1*, and *HES1* displayed significant expression shifts in normal vs. precancerous and normal vs. cancer groups but not between precancerous and cancer tissues, indicating their stable association with esophageal tissue pathology but inability to discriminate malignancy levels. Conversely, *TP53* showed differential expression only between normal and cancer groups, but not between premalignant and normal/cancer samples, consistent with existing knowledge that it is frequently altered in cancers[51], with no evidence linking it to esophageal precancerous lesions.

The gene *NAT8*, which shares N-acetyltransferase activity with *NAT2*, displayed a consistent differential expression pattern between premalignant and normal/cancer samples. Moreover, *NAT8* emerged as a significant predictor of survival in esophageal carcinoma based on survival analysis (Fig. 3E), which demonstrated that PCMR could serve as important resource for mining precancerous prognosis biomarkers. DAVID functional enrichment analysis of 13 genes with significant differential expression in esophageal precancerous lesions revealed strong enrichment in GO biological processes related to dorsal/ventral neural tube patterning and smooth muscle tissue development (Fig. 3F). Neural tube patterning involves differentiation during embryogenesis, mediated by pathways such as Wnt, Notch, and BMP signaling[52–54], all of which are closely linked to cancer initiation and progression. Additionally, aberrant activation of genes involved in esophageal smooth muscle development may indicate early neoplastic transformation[55–57]. These findings demonstrate the reliability and biological relevance of PCMR dataset for studying precancerous lesions.

Usage Notes

In addition to Figshare[36], the data associated with this work is also available at http://www.bio-data.cn/pcmr, which allows for interactive visualization of PCMR datasets.

Code availability

The analysis code is available at https://github.com/ZhoulabCPH/PCMR.

Received: 12 December 2024; Accepted: 25 March 2025;
Published online: 01 April 2025

References

1. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **74**, 229–263, https://doi.org/10.3322/caac.21834 (2024).
2. Chang, J. *et al.* Genomic alterations driving precancerous to cancerous lesions in esophageal cancer development. *Cancer cell* **41**, 2038–2050 e2035, https://doi.org/10.1016/j.ccell.2023.11.003 (2023).
3. Faubert, B., Solmonson, A. & DeBerardinis, R. J. Metabolic reprogramming and cancer progression. *Science* **368**, https://doi.org/10.1126/science.aaw5473 (2020).
4. Ushijima, T., Clark, S. J. & Tan, P. Mapping genomic and epigenomic evolution in cancer ecosystems. *Science* **373**, 1474–1479, https://doi.org/10.1126/science.abh1645 (2021).
5. Menakuru, S. R., Brown, N. J., Staton, C. A. & Reed, M. W. Angiogenesis in pre-malignant conditions. *British journal of cancer* **99**, 1961–1966, https://doi.org/10.1038/sj.bjc.6604733 (2008).
6. Mehrotra, R., Gupta, A., Singh, M. & Ibrahim, R. Application of cytology and molecular biology in diagnosing premalignant or malignant oral lesions. *Molecular cancer* **5**, 11, https://doi.org/10.1186/1476-4598-5-11 (2006).
7. Prime, S. S., Cirillo, N., Cheong, S. C., Prime, M. S. & Parkinson, E. K. Targeting the genetic landscape of oral potentially malignant disorders has the potential as a preventative strategy in oral cancer. *Cancer letters* **518**, 102–114, https://doi.org/10.1016/j.canlet.2021.05.025 (2021).
8. Koop, H. Gastroesophageal reflux disease and Barrett's esophagus. *Endoscopy* **34**, 97–103, https://doi.org/10.1055/s-2002-19851 (2002).
9. Sethi, N. S. *et al.* Early TP53 alterations engage environmental exposures to promote gastric premalignancy in an integrative mouse model. *Nature genetics* **52**, 219–230, https://doi.org/10.1038/s41588-019-0574-9 (2020).
10. Spira, A. *et al.* Leveraging premalignant biology for immune-based cancer prevention. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 10750–10758, https://doi.org/10.1073/pnas.1608077113 (2016).
11. Kang, T. W. *et al.* Senescence surveillance of pre-malignant hepatocytes limits liver cancer development. *Nature* **479**, 547–551, https://doi.org/10.1038/nature10599 (2011).
12. Beane, J. E. *et al.* Molecular subtyping reveals immune alterations associated with progression of bronchial premalignant lesions. *Nature communications* **10**, 1856, https://doi.org/10.1038/s41467-019-09834-2 (2019).
13. Mayinger, B. *et al.* Early detection of premalignant conditions in the colon by fluorescence endoscopy using local sensitization with hexaminolevulinate. *Endoscopy* **40**, 106–109, https://doi.org/10.1055/s-2007-967019 (2008).

14. Hackshaw, A., Clarke, C. A. & Hartman, A. R. New genomic technologies for multi-cancer early detection: Rethinking the scope of cancer screening. *Cancer cell* **40**, 109–113, https://doi.org/10.1016/j.ccell.2022.01.012 (2022).
15. Schrag, D. *et al*. Blood-based tests for multicancer early detection (PATHFINDER): a prospective cohort study. *Lancet* **402**, 1251–1260, https://doi.org/10.1016/S0140-6736(23)01700-2 (2023).
16. Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M. & Wallace, M. B. Colorectal cancer. *Lancet* **394**, 1467–1480, https://doi.org/10.1016/S0140-6736(19)32319-0 (2019).
17. Fitzgerald, R. C., Antoniou, A. C., Fruk, L. & Rosenfeld, N. The future of early cancer detection. *Nature medicine* **28**, 666–677, https://doi.org/10.1038/s41591-022-01746-x (2022).
18. Oeffinger, K. C. *et al*. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. *Jama* **314**, 1599–1614, https://doi.org/10.1001/jama.2015.12783 (2015).
19. Crosby, D. *et al*. Early detection of cancer. *Science* **375**, eaay9040, https://doi.org/10.1126/science.aay9040 (2022).
20. Curtius, K., Wright, N. A. & Graham, T. A. Evolution of Premalignant Disease. *Cold Spring Harbor perspectives in medicine* **7**, https://doi.org/10.1101/cshperspect.a026542 (2017).
21. Singhi, A. D. & Wood, L. D. Early detection of pancreatic cancer using DNA-based molecular approaches. *Nature reviews. Gastroenterology & hepatology* **18**, 457–468, https://doi.org/10.1038/s41575-021-00470-0 (2021).
22. Campbell, J. D. *et al*. The Case for a Pre-Cancer Genome Atlas (PCGA). *Cancer prevention research* **9**, 119–124, https://doi.org/10.1158/1940-6207.CAPR-16-0024 (2016).
23. Zhou, R., Tang, X. & Wang, Y. Emerging strategies to investigate the biology of early cancer. *Nature reviews. Cancer* **24**, 850–866, https://doi.org/10.1038/s41568-024-00754-y (2024).
24. Clough, E. & Barrett, T. The Gene Expression Omnibus Database. *Methods in molecular biology* **1418**, 93–110, https://doi.org/10.1007/978-1-4939-3578-9_5 (2016).
25. Cancer Genome Atlas Research, N. *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* **45**, 1113–1120, https://doi.org/10.1038/ng.2764 (2013).
26. Tomczak, K., Czerwinska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* **19**, A68–77, https://doi.org/10.5114/wo.2014.47136 (2015).
27. Zhang, J. *et al*. The International Cancer Genome Consortium Data Portal. *Nature biotechnology* **37**, 367–369, https://doi.org/10.1038/s41587-019-0055-9 (2019).
28. Zehir, A. *et al*. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine* **23**, 703–713, https://doi.org/10.1038/nm.4333 (2017).
29. Cerami, E. *et al*. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2**, 401–404, https://doi.org/10.1158/2159-8290.CD-12-0095 (2012).
30. Ritchie, M. E. *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47, https://doi.org/10.1093/nar/gkv007 (2015).
31. Sondka, Z. *et al*. COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic acids research* **52**, D1210–D1217, https://doi.org/10.1093/nar/gkad986 (2024).
32. Repana, D. *et al*. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome biology* **20**, 1, https://doi.org/10.1186/s13059-018-1612-0 (2019).
33. Abbott, K. L. *et al*. The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic acids research* **43**, D844–848, https://doi.org/10.1093/nar/gku770 (2015).
34. Jiang, Q. *et al*. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic acids research* **37**, D98–104, https://doi.org/10.1093/nar/gkn714 (2009).
35. Fan, C. *et al*. CircR2Disease v2.0: An Updated Web Server for Experimentally Validated circRNA-disease Associations and Its Application. *Genomics, proteomics & bioinformatics* **20**, 435–445, https://doi.org/10.1016/j.gpb.2021.10.002 (2022).
36. Xiong, Y. *et al*. PCMR, A Comprehensive PreCancerous Molecular Repository and Online Analysis Platform. *figshare* https://doi.org/10.6084/m9.figshare.27997619 (2025).
37. Weaver, J. M., Ross-Innes, C. S. & Fitzgerald, R. C. The '-omics' revolution and oesophageal adenocarcinoma. *Nature reviews. Gastroenterology & hepatology* **11**, 19–27, https://doi.org/10.1038/nrgastro.2013.150 (2014).
38. Alvi, M. A. *et al*. DNA methylation as an adjunct to histopathology to detect prevalent, inconspicuous dysplasia and early-stage neoplasia in Barrett's esophagus. *Clinical cancer research: an official journal of the American Association for Cancer Research* **19**, 878–888, https://doi.org/10.1158/1078-0432.CCR-12-2880 (2013).
39. Liu, S. P. *et al*. LAMP2 as a Biomarker Related to Prognosis and Immune Infiltration in Esophageal Cancer and Other Cancers: A Comprehensive Pan-Cancer Analysis. *Frontiers in oncology* **12**, 884448, https://doi.org/10.3389/fonc.2022.884448 (2022).
40. Kim, S. M. *et al*. Prognostic biomarkers for esophageal adenocarcinoma identified by analysis of tumor transcriptome. *PloS one* **5**, e15074, https://doi.org/10.1371/journal.pone.0015074 (2010).
41. Gharahkhani, P. *et al*. Genome-wide association studies in oesophageal adenocarcinoma and Barrett's oesophagus: a large-scale meta-analysis. *The Lancet. Oncology* **17**, 1363–1373, https://doi.org/10.1016/S1470-2045(16)30240-6 (2016).
42. Callahan, Z. M., Shi, Z., Su, B., Xu, J. & Ujiki, M. Genetic variants in Barrett's esophagus and esophageal adenocarcinoma: a literature review. *Diseases of the esophagus: official journal of the International Society for Diseases of the Esophagus* **32**, https://doi.org/10.1093/dote/doz017 (2019).
43. Malik, M. A., Upadhyay, R., Modi, D. R., Zargar, S. A. & Mittal, B. Association of NAT2 gene polymorphisms with susceptibility to esophageal and gastric cancers in the Kashmir Valley. *Archives of medical research* **40**, 416–423, https://doi.org/10.1016/j.arcmed.2009.06.009 (2009).
44. Matejcic, M. & Iqbal Parker, M. Gene-environment interactions in esophageal cancer. *Critical reviews in clinical laboratory sciences* **52**, 211–231, https://doi.org/10.3109/10408363.2015.1020358 (2015).
45. Yang, L. *et al*. DNA of neutrophil extracellular traps promotes cancer metastasis via CCDC25. *Nature* **583**, 133–138, https://doi.org/10.1038/s41586-020-2394-6 (2020).
46. Zhou, S. *et al*. The ABC transporter Bcrp1/ABCG2 is expressed in a wide variety of stem cells and is a molecular determinant of the side-population phenotype. *Nature medicine* **7**, 1028–1034, https://doi.org/10.1038/nm0901-1028 (2001).
47. Huang, L. *et al*. ABCG2/V-ATPase was associated with the drug resistance and tumor metastasis of esophageal squamous cancer cells. *Diagnostic pathology* **7**, 180, https://doi.org/10.1186/1746-1596-7-180 (2012).
48. Zhang, M. *et al*. Mithramycin represses basal and cigarette smoke-induced expression of ABCG2 and inhibits stem cell signaling in lung and esophageal cancer cells. *Cancer research* **72**, 4178–4192, https://doi.org/10.1158/0008-5472.CAN-11-3983 (2012).
49. Yang, W. & Yu, J. Immunologic function of dendritic cells in esophageal cancer. *Digestive diseases and sciences* **53**, 1739–1746, https://doi.org/10.1007/s10620-007-0095-8 (2008).
50. O'Neill, J. R. *et al*. Multi-Omic Analysis of Esophageal Adenocarcinoma Uncovers Candidate Therapeutic Targets and Cancer-Selective Posttranscriptional Regulation. *Molecular & cellular proteomics: MCP* **23**, 100764, https://doi.org/10.1016/j.mcpro.2024.100764 (2024).
51. Donehower, L. A. *et al*. Integrated Analysis of TP53 Gene and Pathway Alterations in The Cancer Genome Atlas. *Cell reports* **28**, 3010, https://doi.org/10.1016/j.celrep.2019.08.061 (2019).
52. Stolfi, A., Wagner, E., Taliaferro, J. M., Chou, S. & Levine, M. Neural tube patterning by Ephrin, FGF and Notch signaling relays. *Development* **138**, 5429–5439, https://doi.org/10.1242/dev.072108 (2011).

53. Chesnutt, C., Burrus, L. W., Brown, A. M. & Niswander, L. Coordinate regulation of neural tube patterning and proliferation by TGFbeta and WNT activity. *Developmental biology* **274**, 334–347, https://doi.org/10.1016/j.ydbio.2004.07.019 (2004).
54. Timmer, J. R., Wang, C. & Niswander, L. BMP signaling patterns the dorsal and intermediate neural tube via regulation of homeobox and helix-loop-helix transcription factors. *Development* **129**, 2459–2472, https://doi.org/10.1242/dev.129.10.2459 (2002).
55. McCarthy, A. J. & Chetty, R. Benign Smooth Muscle Tumors (Leiomyomas) of Deep Somatic Soft Tissue. *Sarcoma* **2018**, 2071394, https://doi.org/10.1155/2018/2071394 (2018).
56. Weiss, S. W. Smooth muscle tumors of soft tissue. *Advances in anatomic pathology* **9**, 351–359, https://doi.org/10.1097/00125480-200211000-00004 (2002).
57. Domansk, H. A. & Walther, C. S. Smooth-Muscle Tumors. *Monographs in clinical cytology* **22**, 64–67, https://doi.org/10.1159/000475096 (2017).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (No. 62372331). The funders had no roles in study design, data collection and analysis, publication decision, or manuscript preparation.

Author contributions

M.Z. and J.S. contributed to conception and design; Y.C.X., J.Q.L., W.J., X.R.S., H.P., Z.Y.W., C.F.J., L.L.Z., J.Z.H., M.D.Z. and B.B.L. contributed to data collection, analysis and validation. Y.B.Z. contributed to text mining. Y.C.X. constructed the website. Y.C.X., M.Z. and J.S. drafted and revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.S. or M.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.